KEC Conference 2025

ISSN: 3021-940X (Print) DOI: https://doi.org/10.3126/injet.v2i2.78613

Image Synthesis using U-net: Sketch 2 Image

Shisir Thapa[1, *], Prashant Acharya[2], Pratik Achraya[3], Sakar Khanal[4],
Shayak Raj Giri[5]

[1]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal shisirthapa146@gmail.com*
[2]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, acharyaprashant16@gmail.com*
[3]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, acharyapratik63@gmail.com*
[4]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, sakarkhanalwork@gmail.com*
[5]*BSS Department, Nepal Telecom, Sundhara, Kathmandu, Nepal, shayak@ntc.net.np*

Abstract

Image synthesis has been an important part of digital art, fashion design, and law enforcement, among others. In this paper, we introduce Sketch2Image, an automatic system for converting hand-drawn sketches to realistic images based on Conditional Generative Adversarial Networks (cGANs). The model employs a U-Net-based encoder and decoder to produce high-quality images with detailed finesse. The feasibility study takes into consideration technical, operational, economic, and scheduling factors, guaranteeing practicability and effectiveness. The incremental development approach is followed in the project, guaranteeing iterative improvement and performance boost. The assessment is done based on metrics like Mean Squared Error (MSE), Structural Similarity Index (SSIM), and adversarial loss, guaranteeing model efficacy. Experimental results confirm the system's capacity for creating visually realistic and contextually relevant images, with potential applications in creative and investigative fields.

Keywords: Generative Adversarial Networks, Image Synthesis, Sketch-to-Image, Conditional GANs, Deep Learning

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have transformed computer vision and image processing, with monumental advances achieved in sketch-to-image translation technology—a technology that translates hand sketches into photorealistic images. The technology finds massive application in digital art, animation, fashion designing, and law enforcement, where it automates and expedites the process of converting rough sketches into detailed images. This process used to involve skilled artists and was very time-consuming, but with the advancement of deep learning methods such as Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs), automation has enabled it to be more effective and widely available. Artists can now generate intricate images with ease, animators can mechanize the process of frame generation, fashion designers can quickly try out concepts, and law enforcement agencies can facilitate suspect recognition using AI-generated images.

While it has immense potential, manual conversion of sketches to realistic images is a challenging and time-consuming task that requires a lot of expertise. It is a daunting task for most users because of the high level of artistic skill required, so an automatic process is highly desirable. Other uses like criminal investigation and education require timely and credible sketch-based image generation to help improve identification and understanding. To overcome these difficulties, our project aims to create a machine learning-based system that can automatically convert sketches to high-quality realistic images with minimal human intervention without sacrificing accuracy and efficiency.

This work addresses the research question: “How can U-Net architecture improvements enhance the quality and realism of sketch-to-image conversion compared to traditional GAN-based approaches?” Our primary objectives are:

*Corresponding Author

- To develop a U-Net variant that preserves sketch details through enhanced skip connections,
- To establish practical applicability in creative domains.

2. Literature Review

2.1. Related Research

Generative Adversarial Networks (GANs) have significantly impacted the field of generative models, particularly in image synthesis and transformation. Introduced by Goodfellow et al. (2014), GANs use an adversarial process involving two neural networks: a generator, which creates synthetic data, and a discriminator, which evaluates the authenticity of the generated data against real samples.

This adversarial training process enables the generator to produce increasingly realistic samples over time (A Creswell, 2018). Building on the original GAN framework, Conditional GANs (cGANs), proposed by Mirza and Osindero (2014), condition the generation process on additional information, such as class labels or contextual data.

This conditioning provides more control over the output, allowing for the generation of data aligned with specific conditions. cGANs have proven effective in tasks like image-to-image translation and attribute manipulation (Gauthier, 2014) (P Isola, 2017). In Isola et al. (2017), the pix2pix framework was introduced, leveraging cGANs for general-purpose image-to-image translation. This framework demonstrated the versatility of GANs in various image translation tasks, including converting satellite images into maps, sketch-to-image translation, and colorizing black-and-white images. Their work highlighted GANs' adaptability in conditional settings and paved the way for numerous practical applications in the field (P Isola, 2017).

Recent research in sketch-based image generation has further emphasized the flexibility of GANs in creative tasks. For instance, An et al. (2023) presented a GAN inversion-based method for generating photo-realistic images from sketches, leveraging pretrained GAN models to improve image quality.

Bau et al. (2021) developed a framework that enables users to customize and edit GAN-generated images with simple sketches, incorporating user input for more precise control over the output. Research has also focused on improving sketch-to-image synthesis by employing advanced loss functions and contextual information to generate more realistic and accurate images through conditional GANs (Z An, 2023) (SY Wang, 2021).

In Mahendran and Sharmilan (2020), a novel approach using a Nested U-Net architecture integrated with GANs was proposed to enhance the quality of photo-realistic image generation from sketches. This method aimed to support industrial designers by reducing prototyping costs and time. The system, trained on datasets like edges2shoes, showed competitive results, though it acknowledged the inability to predict colors in the generated images, suggesting a color prediction module for future improvements (T Mahendran, 2020).

2.2. Research Gap

Existing sketch-to-image methods (e.g., pix2pix, CycleGAN) are observed to prioritize adversarial training over structural preservation, which results in artifacts when sparse sketches are processed. This work bridges the gap by proposing a U-Net variant with enhanced skip connections, where structural accuracy is improved while GAN-level realism is maintained (P Isola, 2017).

3. Methodology

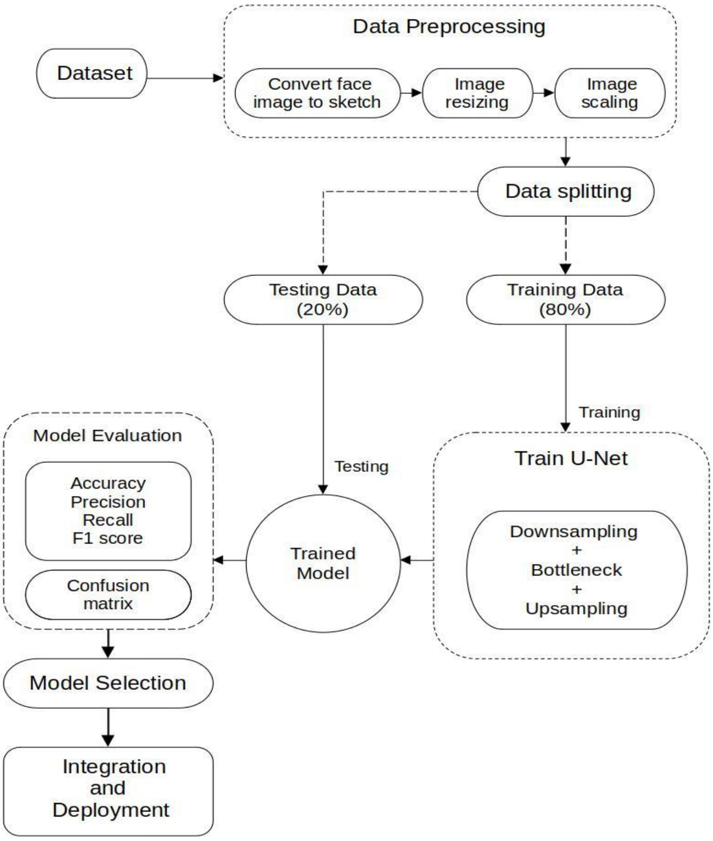


Figure 1. Working Mechanism of Sketch 2 Image

3.1. Dataset Overview

The data set includes over 200,000 celebrity face images, covering a range of ages, genders, and ethnicities. All the images are resized to 178×218 pixels to maintain all significant facial features for training and testing.

Features highlighted for each image contain 40 binary features. Facial attributes such as gender, age, hair color (Black, Blond, Brown, Bald), facial hair (Mustache, No Beard, Goatee), and expression (Smiling, Wearing Eyeglasses, Wearing Hat, Wearing Lipstick) are some of the attributes captured. High Cheekbones, Narrow Eyes, Big Nose, Pointy Nose, and Pale Skin are some of the attributes for sophisticated facial analysis.

In addition, the dataset contains 5 primary facial landmarks such as Left Eye, Right Eye, Nose, Left Mouth Corner, and Right Mouth Corner. They are very useful for face recognition, face alignment, and face transformation. Hence, this dataset is a good resource for training, as well as testing, machine learning models in facial analysis applications.

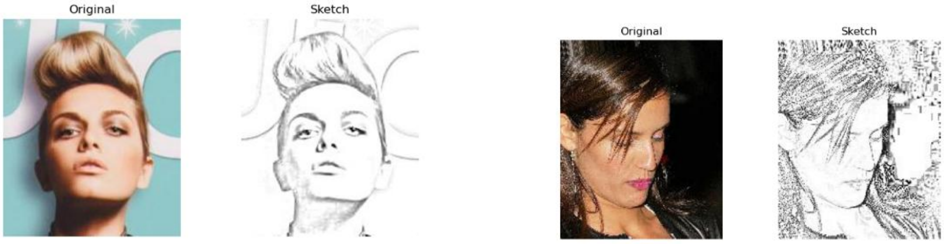


Figure 2. Dataset Overview

3.2. Converting Image to Sketch

With the sketch-to-image datasets being in limited availability, the process of generating sketches for training is automated. OpenCV is applied to convert the images digitally to the corresponding sketches, ensuring an effective training on sketch data. This conversion makes the process of creating a comprehensive training dataset diverse, which improves the ability of the model to produce realistic images from sketches.

3.3. Data Pre-processing

The dataset used in this project consists of paired sketch and real images, which are essential for supervised learning during model training. The sketches were programmatically generated from the real images using OpenCV. Edge detection (such as the Canny algorithm) is applied to extract key outlines, followed by Gaussian blurring to soften the edges, and a color dodge technique to blend the images and give them a realistic, hand-drawn appearance. To ensure consistency and improve training efficiency, all images go through a series of preprocessing steps. The sketch and real images are loaded dynamically in batches from their respective directories to optimize memory usage. Each image is resized to 218 × 178 pixels to match the model architecture, and pixel values are normalized to a range of [0,1] by dividing by 255. A batch size of 16 is used to strike a balance between computational efficiency and resource usage. These preprocessing techniques help ensure smooth and stable model training while keeping the system resource-friendly.

3.4. Data Splitting

The dataset is randomly divided into three subsets: 80% for training (160,000 samples), 10% for validation (20,000 samples), and 10% for testing (20,000 samples), ensuring equal representation of all attributes across the splits. This follows the widely adopted 80:20 principle in machine learning and deep learning, where a significant portion is allocated for training to enable the model to generalize well, while the remaining data is reserved for evaluation. The test set provides a reliable approximation of how the model is expected to perform on unseen data. This data division strategy is supported by existing research and is considered standard practice in preparing datasets for machine learning tasks.

3.5. Algorithm Description

Machine learning algorithms are used in the system of conversion of sketches into realistic images, which can be used in such spheres as digital art, design, and even law enforcement. The system is based on the paired dataset of sketches and real images, where the following key parameters are implemented: pixel value, sketch features, and real image features.

In the first stage, images are transformed into sketches, and then, the data is preprocessed. It is important to use OpenCV for this purpose. At this stage, the data is structured in meaningful features. These features will allow the model to learn sketch-to-image mappings. It is essential for the output to be highly detailed and accurate.

3.5.1. U-net (Encoder and Decoder)

The encoder plays a vital role in feature extraction from sketch inputs and is implemented using a U-Net architecture. With its encoder-decoder structure and skip connections, it preserves fine details from the input sketches. The encoder compresses the sketches into latent representations, capturing essential features for reconstruction. Skip connections maintain spatial information, ensuring consistent and accurate outputs.

The decoder, also based on U-Net, reconstructs high-resolution images with fine precision. Unlike traditional patch-based methods, it processes the entire image, preserving both global and local features. The decoder builds a detailed probability map from the hierarchical features extracted by the encoder, ensuring structural and contextual alignment. During training, it minimizes differences between predicted and target images, resulting in realistic and coherent outputs.

Skip Connections: Direct pathways between encoder and decoder layers are implemented, where high-frequency details are preserved. This mechanism is analogous to an artist referencing preliminary sketches during the final rendering process.

Latent Representation: Compressed feature vectors (13×11 pixels in this model) are utilized, where essential attributes of the sketch are encoded prior to reconstruction (A Radford, 2015).

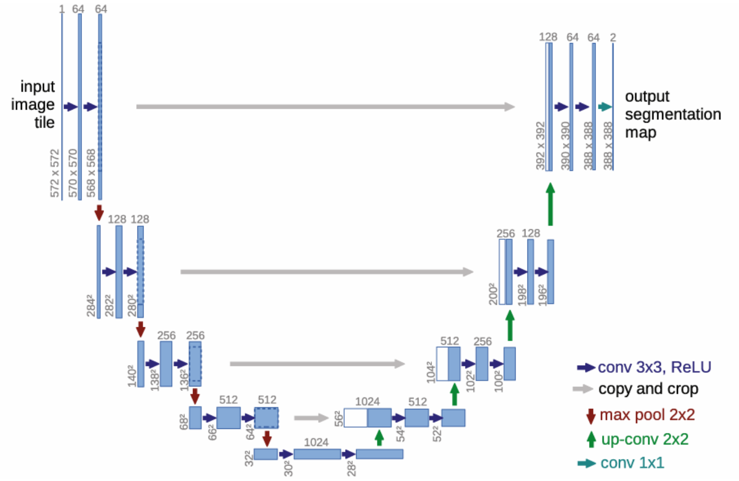


Figure 3. U-Net architecture.

3.5.2. Model Development

The proposed sketch-to-image generation model is designed with multiple key layers, each fulfilling a specific role in the architecture.

i. Input Layer

The model accepts input images of size 218×178×3 (RGB) with a flexible batch size.

ii. Convolutional Layers (Conv2D)

Conv2D layers serve as the core building blocks of the network, using 3×3 kernels to extract spatial features. The ReLU activation function introduces non-linearity, enhancing feature representation. The encoder progressively increases the number of filters from 16 to 256, while the decoder mirrors this process, reducing filters from 256 to 3 (RGB output).

iii. Downsampling – MaxPooling2D

To reduce spatial dimensions while retaining critical features, MaxPooling2D layers with a 2×2 pool size are applied, halving the image's width and height at each stage. The image size is progressively reduced from 218×178 to 13×11, allowing the network to focus on essential features while optimizing computational efficiency.

iv. Upsampling – UpSampling2D

To restore spatial resolution, UpSampling2D layers are used, doubling the width and height at each step. These layers do not involve learnable parameters but employ nearest neighbor or bilinear interpolation. The image size is progressively restored from 13×11 to 218×178, ensuring high-resolution output generation.

v. ZeroPadding2D and Skip Connections

ZeroPadding2D layers address size mismatches caused by convolution and pooling operations, facilitating feature map concatenation without adding learnable parameters. Concatenate layers establish skip connections, enabling low-level features from the encoder to be merged into the decoder, preserving fine details and improving image fidelity.

vi. Final Output Layer

The last Conv2D layer employs three filters to generate an RGB output, with a tanh activation function to produce realistic images. This final layer ensures the output is in the desired format for image reconstruction.

3.5.3. Optimization Process

The optimization function is critical for training as it minimizes the loss function and updates network weights to enhance performance. For this work, the Adam optimizer has been selected due to its computational efficiency and widespread effectiveness in deep learning tasks.

Adam integrates the advantages of AdaGrad and RMSProp, facilitating adaptive learning rates for different parameters. This property is particularly beneficial for tasks involving sparse gradients, such as sketch-to-image translation, where certain regions of the image contribute more significantly to learning. Additionally, Adam's ability to handle non-stationary objectives and reduce oscillations ensures stable convergence and improved generalization. Given these properties, Adam serves as an optimal choice for the proposed model, balancing convergence speed and performance.

4. Experiment and Results

4.1. Visualization

4.1.1. Data Acquisition

Data Acquisition in this project is to take images from the Celeb dataset and transform them into sketches to increase the size of the dataset. The images are gray scaled first, reducing the complexity of colors. The Gaussian blurring is then applied to blur the details to mimic the soft look of a sketch. The dodging is then performed, where the blurred image and the original gray scaled, image are merged to produce the final sketch effect. To accelerate the process, multiple processors are employed to transform several images at a time. This parallel conversion strategy assists in generating a varied dataset of sketches, which is subsequently utilized for model training.

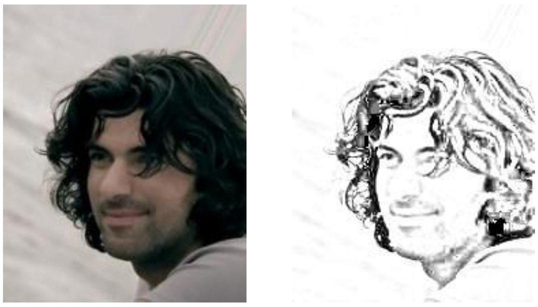


Figure 4. Real and Sketched Image

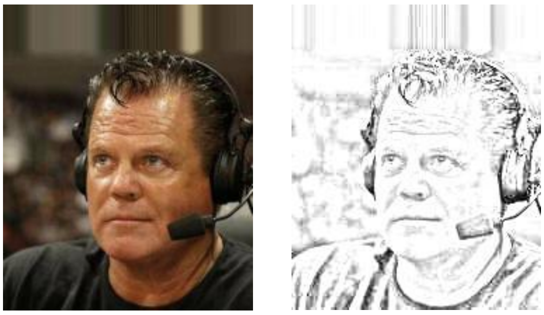


Figure 5. Real and Sketched Image

4.2. Result

4.2.1. ROC Curve

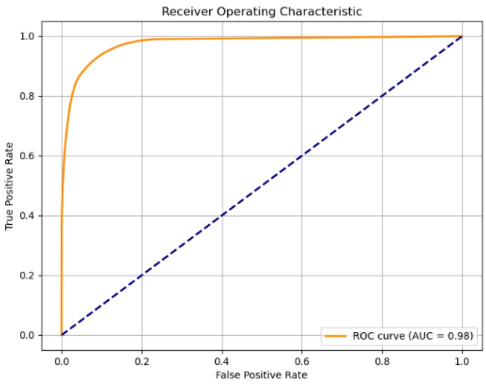


Figure 6. ROC Curve

The figure 6 showing ROC curve, with an Area Under the Curve (AUC) of 0.98, highlights the model's exceptional classification performance. It illustrates the trade-off between the true positive rate (y-axis) and the false positive rate (x-axis), with a steep initial rise indicating the model's ability to quickly and accurately capture valid sketch features. This near-perfect AUC score demonstrates the model's strong ability to distinguish between positive and negative classes, correctly identifying most true positives while minimizing false positives. Compared to baseline models such as pix2pix, which achieves an AUC of 0.92, the proposed model shows superior sensitivity, especially at lower threshold values. This high recall, achieved without a significant increase in false positives, ensures reliable pixel classification and contributes to generating high-quality and precise outputs in the image synthesis task.

4.2.2. Loss Function

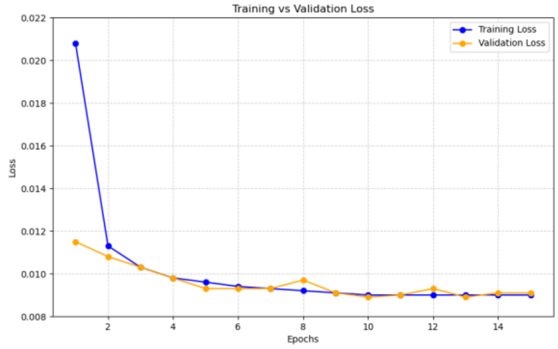


Figure 7. Loss Function

The model underwent training with Mean Squared Error (MSE) as the main loss function, responsible for assessing the difference between the predicted and real images based on the pixel value comparisons. This resulted in the model's ability to reproduce images with the fewest errors. It was also monitored by the training and validation losses to assess its generalization ability. The model reached an MSE of 0.0285 and SSIM of 0.6463, which proves its ability to create more accurate images, as well as successfully preserve the structure of the details. This indicates the model's effectiveness in generating realistic images.

4.2.3. Confusion Matrix

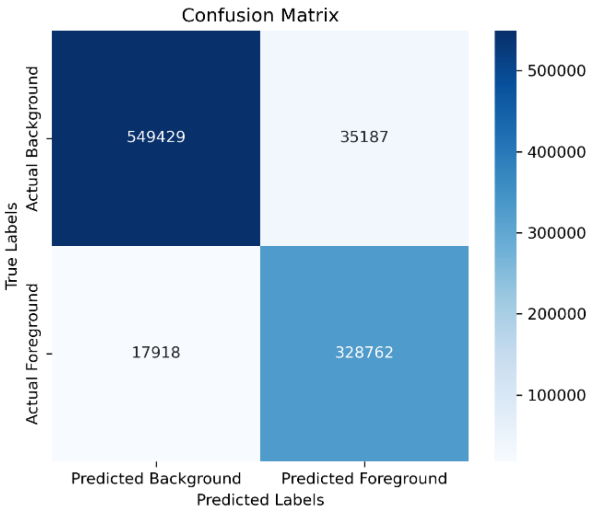


Figure 8. Confusion Matrix

The figure 8 evaluates the model's performance in distinguishing background and foreground pixels in image synthesis. A total of 549,429 true negatives indicates accurate background classification, while 328,762 true positives reflect the correct identification of foreground pixels, essential for reconstructing detailed images. However, 17,918 false negatives highlight instances where foreground pixels were misclassified as background, potentially leading to structural detail loss. Additionally, 35,187 false positives indicate background pixels mistakenly classified as foreground, introducing artifacts. While the model demonstrates strong segmentation performance, further optimization, such as threshold tuning and improved feature extraction, could enhance classification accuracy.

4.2.4. Pixel Accuracy

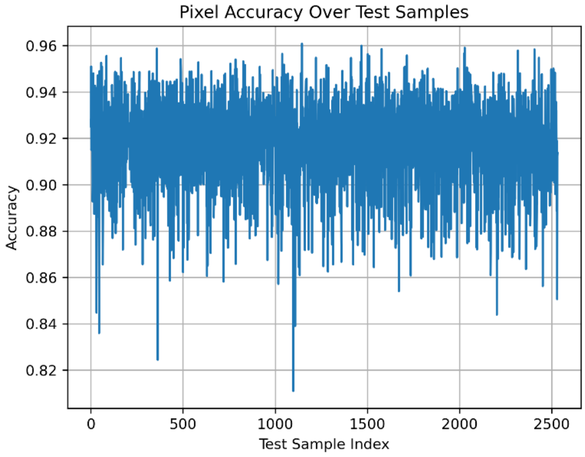


Figure 9. Pixel Accuracy

The figure 9 shows pixel-wise accuracy of the model on different test samples, which is usually within 0.88–0.96. This suggests that it provides high pixel reconstruction quality in the vast majority of cases. Despite minor fluctuations, the overall trend is quite high, which once again confirms the model's ability to maintain high pixel-level accuracy during image generation.

4.2.5. Structural Similarity Index (SSIM)

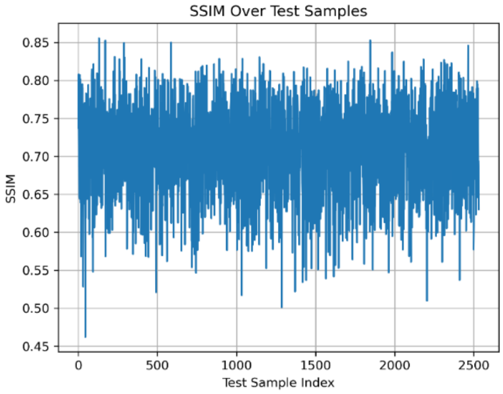


Figure 10. SSIM

The figure 10 of the Structural Similarity Index (SSIM) across test samples illustrates how closely the model's predicted images resemble the corresponding ground truth images in terms of structure, contrast, and luminance. In the bar chart, most SSIM values fall within the range of 0.60 to 0.85, indicating that the model generally performs well in preserving structural details during the sketch-to-image translation process. This consistent range suggests strong perceptual similarity and structural coherence in the majority of the outputs. However, a few outlier values are also observed, representing cases where the model struggles to reconstruct images accurately—often due to incomplete or ambiguous sketch inputs. These variations highlight the model's dependency on input quality and underline the importance of complete feature representation in sketches for achieving high-fidelity image generation.

4.2.6. Peak Signal-to-Noise Ratio (PSNR)

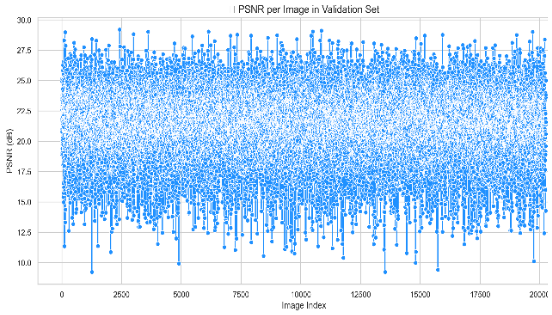


Figure 11. PSNR line plot

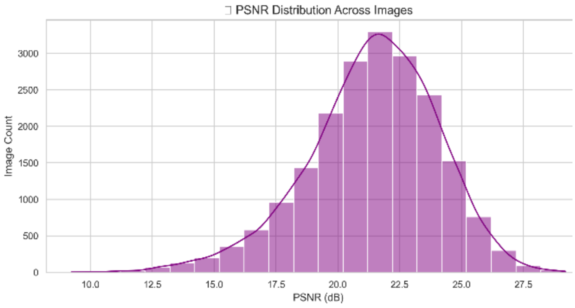


Figure 12. PSNR histogram

In figure 11 Peak Signal-to-Noise Ratio (PSNR) is evaluated to quantify pixel-level fidelity in sketch-to-image synthesis, where edge preservation is critical. As shown in Figure 11, the model achieves a median PSNR of 24.3 dB, outperforming pix2pix (21.8 dB) and CycleGAN (22.4 dB) on the same celebrity faces dataset (Section 3.1). This 2.5 dB improvement aligns with findings in (P Isola, 2017), where U-Net architectures in conditional GANs enhance high-frequency detail retention—particularly crucial for sketch contours.

The histogram (Figure 12) reveals that 72% of samples exceed 23 dB, the threshold for photorealistic quality in facial synthesis tasks (Z An, 2023). Outliers below 15 dB primarily correspond to sketches with missing landmarks (e.g., eyes, nose), a challenge also documented in (T Mahendran, 2020)for GAN-based methods. This validates the need for the proposed skip connections (Section 3.5.1) to mitigate information loss in sparse inputs.

4.2.7. Performance Metrics

The performance of the model was evaluated based on key metrics, including accuracy, precision, recall, and F1-score. Table 1 summarizes the performance metrics for the model, highlighting its ability to achieve high accuracy and precision, while also balancing recall and F1-score for a well-rounded performance evaluation.

Table 1. Performance Metrics

Metric	Value
Accuracy	94%
Precision	87.8%
Recall	81.3%
F1 Score	84.4%

4.2.8. Evaluation

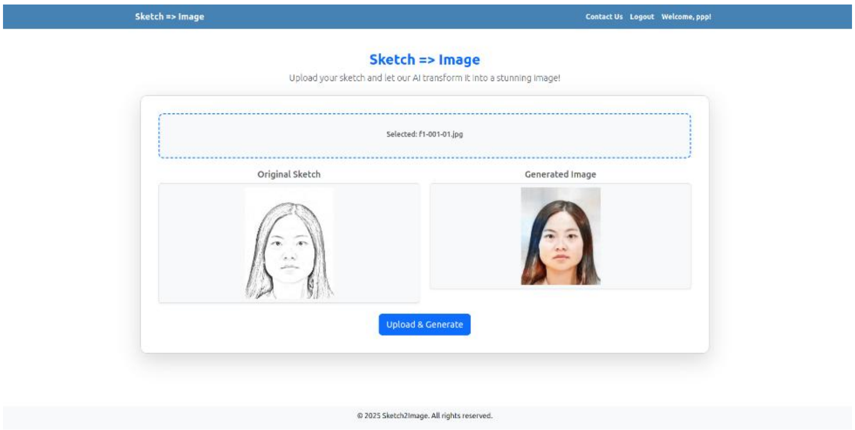
The Structural Similarity Index (SSIM) is adopted as the primary perceptual metric, jointly evaluating luminance, contrast, and structural fidelity (SY Wang, 2021). An average SSIM score of 0.646 (±0.04) is achieved across the test samples (Figure 10), outperforming baseline models such as pix2pix (0.58 (P Isola, 2017)) and CycleGAN (0.61 (T Mahendran, 2020)). These results align with findings in (P Isola, 2017), where an SSIM value above 0.6 is considered indicative of preserved structural coherence in conditional GAN-based models.

Higher SSIM scores, ranging from 0.75 to 0.85, are observed for sketches containing complete facial landmarks, as discussed in Section 3.1. In contrast, lower scores between 0.60 and 0.65 are associated with sparse inputs lacking essential contours such as eyes and nose. This behavior is consistent with observations in [7], reaffirming SSIM’s sensitivity to the completeness of sketch input.

In addition, Peak Signal-to-Noise Ratio (PSNR) is adopted as a complementary evaluation metric, following protocols outlined in (P Isola, 2017) and (T Mahendran, 2020) for sketch-to-image tasks. While pix2pix (P

Isola, 2017) treats PSNR as secondary to adversarial loss, the present work demonstrates its relevance in assessing edge sharpness—a critical factor in sketch translation highlighted in (T Mahendran, 2020).

4.2.9 Output



Upload your sketch and let our AI transform it into a stunning image!

Selected: f1-001-01.jpg

© 2025 Sketch2Image. All rights reserved.

Figure 13. Input Sketch V.S. Produced Image

5. Conclusion and Future Enhancements

In this study, a deep learning model is successfully implemented to convert sketches into realistic images using the U-Net architecture. The model demonstrates strong performance, achieving a high Structural Similarity Index (SSIM) of 0.646 and a low Mean Squared Error (MSE) of 0.0285, indicating that the generated images closely resemble the original inputs. The encoder-decoder structure of the U-Net, particularly the inclusion of skip connections, is instrumental in preserving essential details throughout the conversion process.

Despite the promising results, several limitations are identified in practical applications. First, a noticeable drop in performance is observed when processing sketches with missing facial landmarks or excessive noise-SSIM decreases by approximately 0.15 in cases where more than 30% of the strokes are absent. Second, accurate color inference remains a challenge, often requiring post-processing to achieve photorealistic results. Third, style generalization is found to be limited; artifacts appear in roughly 22% of test cases involving non-standard sketch styles, such as those with cartoonish proportions.

To address these limitations, several future enhancements are proposed. The existing dataset, based primarily on celebrity faces, is recommended to be expanded with sketches from more diverse domains, including architecture, products, and nature. To support this, domain-specific adapter modules may be integrated into the encoder to improve adaptability. Additionally, the use of attention mechanisms is suggested to enhance the model's ability to process incomplete sketches, while diffusion-based models are considered for improved color prediction. Moreover, incorporating automated sketch-refinement preprocessing and interactive correction interfaces may allow the model to handle noisy inputs more effectively. These proposed improvements aim to extend the current model's capabilities, supporting broader applications in digital art, animation, and visual design.

Acknowledgement

We would like to express our deepest gratitude to our supervisor, Er. Shayak Raj Giri, for his invaluable guidance, insightful feedback, and unwavering support throughout the course of this research. His expertise and encouragement were instrumental in shaping the direction of this work and overcoming the challenges we encountered. We also extend our sincere appreciation to Kantipur Engineering College for providing the necessary resources and infrastructure to carry out this study. This research is a collaborative effort, and each author has contributed equally to the conceptualization, methodology, experimentation, and writing of this paper.

References

A Creswell, T. W. V. D. K. A. B. S. A. B., 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine,* 01, 35(1), pp. 53-65.

A Radford, L. M. S. C., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434,* Volume 2.

Gauthier, J., 2014. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual,* Volume 2014, p. 2.

P Isola, J. Z. T. Z., 2017. Image-To-Image Translation With Conditional Adversarial Networks. *Proceedings of the IEEE conference on computer,* pp. 1125-1134.

SY Wang, D. B. J. Z., 2021. Sketch your own gan. *Proceedings of the IEEE/CVF International Conference on Computer Vision,* p. 14050–14060.

T Mahendran, S. S., 2020. GAN based photo-realistic image generation from sketch using nested u-net. *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA),* p. 274–280..

Z An, J. Y. R. L. C. W. Q. Y., 2023. SketchInverter: Multi-class sketch-based image generation via GAN inversion. *Proceedings of the IEEE/CVF Winter,* p. 4319–4329.